Research Report

Scale Drift in Equating on a Test That Employs Cut Scores

Gautam Puhan

Research & Development

July 2007
RR-07-34

Scale Drift in Equating on a Test That Employs Cut Scores

Gautam Puhan

ETS, Princeton, NJ

July 2007

Abstract

The purpose of this study is to determine the extent of scale drift on a test that employs cut scores. It is essential to examine scale drift in a testing program using new forms that are often put on scale through a series of intermediate equatings (known as equating chains). This may cause equating error to accumulate to a point where scale scores are rendered incomparable across two parallel chains or time periods. The study examined whether scale drift occurred for two conditions (i.e., parallel equating chains and a single long chain). Data from three tests that employed cut scores were used in this study. Results indicated that although there was some difference observed between the actual equating conversions derived via different equating chains, the effect of these differences on actual pass/fail status of test takers was not large. However, these results may change if new cut scores are introduced. Also, for tests that do not employ cut scores, stability in most parts of the score scale is important, not just at particular cut score points.

Key words: Scale drift, equating error, equating chains, cut scores, braiding

Acknowledgments

The author would like to thank Mary Grant, Kevin Meara, Dan Eignor, Sandip Sinharay, and Fred McHale for their helpful comments on an earlier draft of the paper and Kim Fryer for editorial assistance.

Table of Contents

Page

List of Tables

List of Figures

Page

Theoretical Background

Test equating is a statistical procedure used to measure and adjust difficulty differences across parallel forms of a test, which in turn allows for score comparisons across different groups of examinees regardless of the test forms they were administered or when they took the test. Test equating involves two general sources of error (Kolen & Brennan, 2004). *Random equating error* is present whenever a sample from a population of examinees is used to estimate the equating relationship. *Systematic equating error* may result from factors such as using a particular equating method when another method is clearly more appropriate, using common items that do not adequately represent the total test in a non-equivalent anchor test (NEAT) design, or employing improper data collection designs. When systematic error is minimized, equating error usually results from sampling variability.

Test equating in actual testing programs often involves a chain of equatings that allow scores to be reported in terms of scaled scores on a common scale. According to Livingston (2004), although equating error is usually not large enough to create problems, it can be a potential problem when the error is consistently in the same direction (i.e., equating error resulting in consistently lower or higher equated scores for successive equatings). Consider a hypothetical equating chain as shown in Figure 1. The equated scores on Form B could be slightly low because of equating error. This may not present an immediate problem for comparing scores on Form B with scores on Form A. Similarly, the equated scores on Form C could be slightly low because of equating error. This may also not present an immediate problem for comparing scores on Form C with scores on Form A. However, if the equating error continues in the same direction for the subsequent equatings for Forms D and E, then the problem can escalate until the equated scores on Form E are no longer comparable to the equated scores on Form A. This is referred to as *scale drift* (Livingston). Since equated scaled scores are intended to be comparable across forms and time, it is essential to monitor the stability of a reporting scale. Standard 4.17 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) emphasizes the need to monitor scale stability. It states that “Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.”

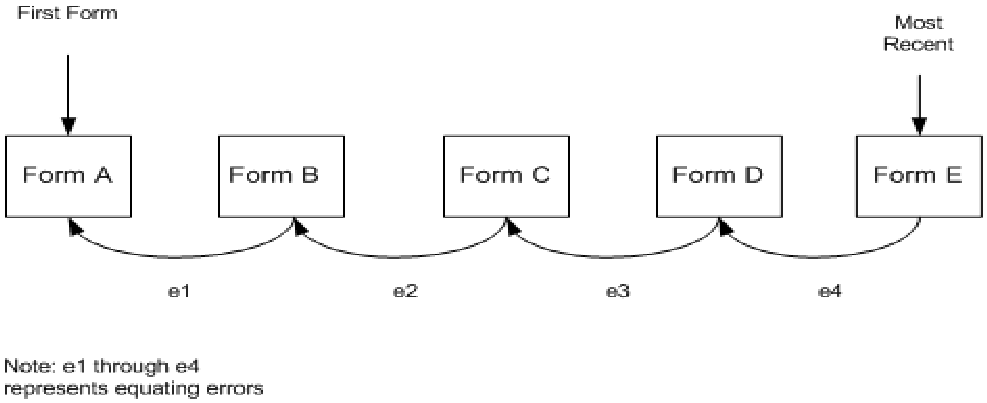

***Figure 1.* Example of a hypothetical equating chain that links to the original base scale.**

Although braiding plans (i.e., an equating plan where tests forms are interweaved to avoid the development of separate "strains") are often used to avoid accumulation of equating error in equating chains, practical considerations, such as not having enough common items to equate a new form back to two or more old forms, often restrict their use. Monitoring scale drift for parallel equating chains or different time periods consequently becomes essential (see Angoff, 1971, chap. 2; Donlon, 1984, chap. 2, for more details on braiding plans).

Previous Studies

A few relevant studies on scale stability are discussed in this section. Stewart (1966) conducted a scale stability study to examine the extent of drift that might have occurred in the Scholastic Aptitude Test (SAT®) scale since 1941 and found that a verbal score of about 500 in 1963 could be 20-35 points higher than the score of a candidate of equal ability tested in 1948. Similarly, a stability study by Modu and Stern (1975) on the SAT scale indicated that the equating scale had drifted by almost 14 points for the verbal section and 17 points for the mathematics section between 1963 and 1973.

However, McHale and Ninneman (1994) assessed the stability of the SAT scale from 1973 to 1984 and found that the SAT-verbal score scale showed little drift. Further, the results from the math scale were inconsistent, and therefore the stability of this scale could not be determined. In a more recent study, Guo and Wang (2003) studied scale drift on a computer adaptive testing (CAT) program and found little evidence for scale drift across two time periods. Nevertheless, the above authors acknowledged the importance of monitoring score scale stability because significant scale drift would likely have a strong impact on test scores and their interpretation.

Petersen, Cook, and Stocking (1983) compared several equating methods in relation to scale drift. Their results indicated that for reasonably parallel tests, linear equating performed well, producing little drift. However when tests varied somewhat in terms of content and length, equating methods based on item response theory (IRT) worked better in maintaining scale stability. Although some of these studies indicate scale stability across different points in time, others clearly show that scale instability can occur. Therefore, it is important to monitor the stability of score scales to ensure that the scores reported to test takers are comparable across different points in time. Moreover, monitoring scale drift is especially important in the case of criterion-referenced tests, where a slight change in test scores can change the pass/fail status of test takers.

The purpose of this study is to determine the extent of scale drift on a test that employs cut scores. Since these tests hold high stakes for test takers, overexposure of items is of constant concern. For this reason, several new forms are developed each year to prevent any particular form from being overexposed. However, this means that each new form gets scaled scores through several intermediate equatings , thereby increasing the potential for scale drift. Although new forms can be equated to one old form to avoid intermediate equatings (e.g., new forms X, Y, and Z equated to old form Q), finding non-overlapping anchor sets for the new forms becomes difficult when several new forms are equated to the same old form. (It is often desirable to have non-overlapping anchor sets for different new forms to ensure that exposure of one new form does not partially expose other new forms.)The study will assess scale stability under two conditions. In the first condition, scale stability will be assessed for two parallel equating chains. In the second condition, scale stability will be assessed for a single long equating chain. Since these tests have cut scores, the effect of scale drift (if observed) on the pass/fail status of examinees will also be assessed.

Method

Test Data

The study used test data from three tests using cut scores. These three tests (named A, B, and C for economy of presentation) had fairly long equating chains (parallel and single) and therefore served as ideal tests for the purpose of this study. Test A consisted of 75 multiplechoice (MC) items assessing basic skills. Test B consisted of 25 constructed-response (CR) items measuring problem-solving skills. Test C consisted of 50 MC items measuring competency in mathematics. (Note that the actual number of items in the test forms analyzed in the current study may differ slightly from the numbers stated above because some items in some forms failed to meet ETS quality standards.)

Analytical Procedure

Equating methods. Depending on the condition (discussed in the next section), either a NEAT design using common items or a randomly equivalent groups (spiraling)[1] design was used to equate new test forms to old test forms. Under the NEAT design, the chained linear and chained equipercentile equating methods were used and under the randomly equivalent groups design, the direct linear and the direct equipercentile equating methods were used. In the chained equating method, new form scores are converted to scores on the common items using examinees who took the new form. Then scores on the old form are converted to scores on the common items using examinees who took the old form. These two conversions are then chained together to produce a conversion of the new form scores to the old form scores (see Livingston, 2004, pp. 43-48, for a detailed discussion of these methods). Following rules of thumb provided by Kolen and Brennan (2004, p. 271), the common items in the NEAT design constituted at least 20% of the length of the total test. They were also chosen to represent the total test in content and difficulty. It should be noted that before the actual equating relationship was derived for the new and old forms, log-linear smoothing (Holland and Thayer, 1987) was applied to the score distributions to adjust for irregularities that may cause problems for the chained or direct equipercentile equating. (Smoothing score distributions does not affect linear equating results.)

Assessing the stability of parallel chains (Condition 1). In this condition, the stability of two parallel chains often referred to as equating strains (as seen in Figure 2) is assessed. Parallel chains may appear when testing programs administer and equate two or more new test forms to one old form, thereby creating separate equating lineages. There is a possible test security reason for creating parallel equating chains. For example, it may be better to equate two new non-overlapping forms (e.g., Forms Y and Z) to a single old form (e.g., Test X) rather than equating Form Y to Form X and Form Z to Form Y. As evident in the second design, if Form Y is exposed, then both Forms X and Z are partially exposed because they share common items with Form Y. But if two separate strains were created and Form Y was exposed, then only Form X would be partially exposed. The purpose was to assess if two parallel chains have drifted enough that equating a new form to an old form on one chain would be different than equating the same new form to an old form in the second chain. Two designs were used to evaluate this. The equating chains shown in Figure 2 represent the two designs. For illustration purposes, hypothetical names for the test forms (e.g., New Form 7) are used.

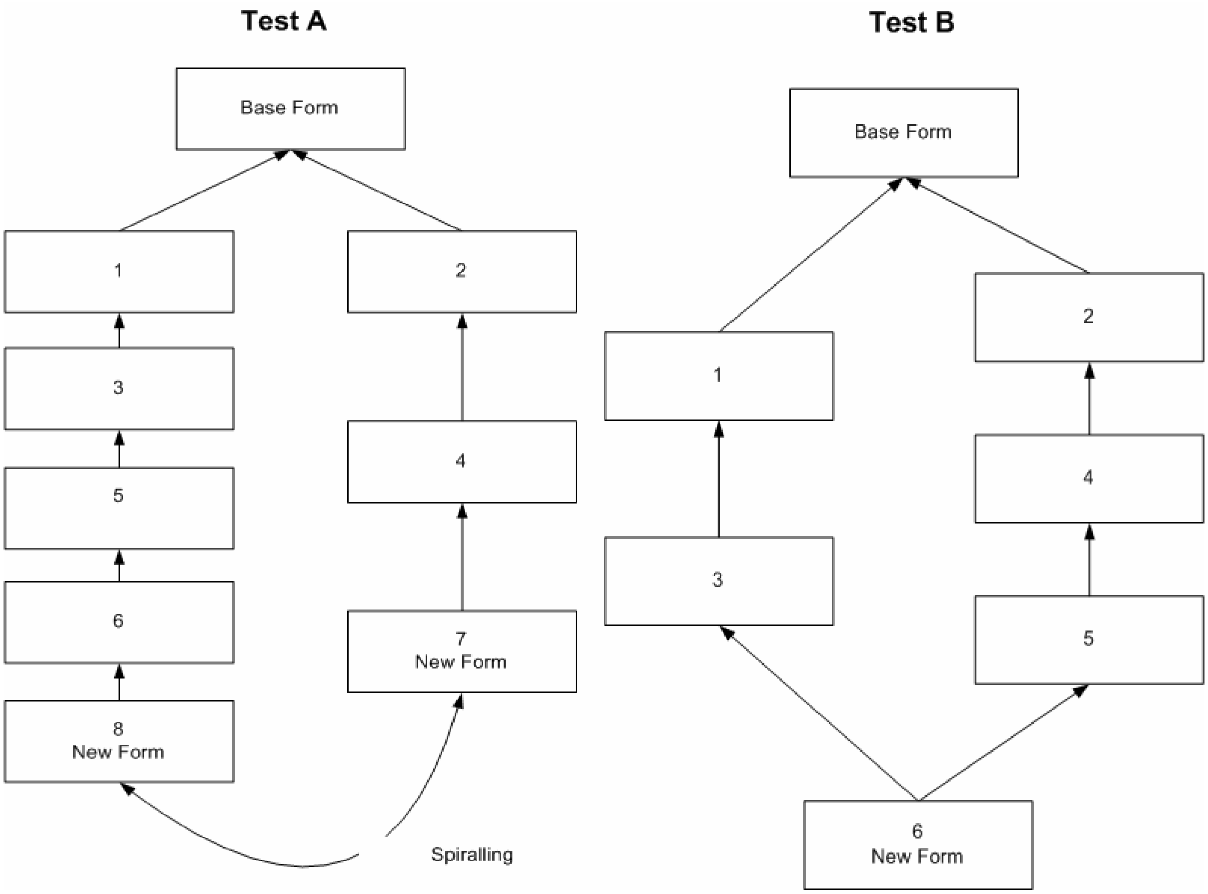

Figure 2. **Parallel chain equating design for Condition 1.**

In the first design, which used Test A (see Figure 2), two new forms were spiraled in a single testing administration. This test had a large volume of test takers and employed the randomly equivalent groups design. As seen in the figure , New Form 7 was first equated to Old Form 4 using common items and was linked to the base form via Form 2, which also used common-item equating. Then New Form 7 was equated to New Form 8 using the randomly equivalent groups design and was subsequently linked to the base form via Forms 6, 5, 3, and 1 using common-item equating. Therefore Form 7 had two equating functions (i.e., one derived from the first equating chain starting with the common-item equating and the other derived from the second equating chain starting with the randomly equivalent groups design). These two equating functions were compared to see if there was a large difference between them, which would indicate a scale drift of the parallel equating chains. Note that a similar analysis could be conducted for New Form 8 that could be linked to the base scale via the two parallel equatings (i.e., the common-item equating and the equating using the randomly equivalent groups design). However, since the analysis of New Form 7 and New Form 8 gave very similar results, it was decided to include only New Form 7 in this study.

In the second design, which used Test B (see Figure 2), New Form 6 was equated to Old Form 5 using a set of common items and then linked to the base form via Forms 4 and 2, which also used common-item equating. Then New Form 6 was equated to Old Form 3 using another set of common items and linked to the base form via Form 1, which also used common-item equating. Both common-item sets used in the two equatings were chosen to be similar in test content and difficulty. The two equating functions were compared to determine whether there was a large difference between them, which would indicate a potential scale drift of the parallel equating chains.

Assessing the stability of a single long chain (Condition 2). As seen in Figure 3, in this design, which used Test C, New Form 4 was equated to the previous old form (i.e., Form 3) using a set of common items. New Form 4 was also equated to another old form used some time ago (i.e., Form 1) using the same set of common items. Note that Forms 1 and 3 had some overlapping items, which made it possible to use the exact same common items in the two equatings. These two equating functions were compared to determine whether there was a large difference between them, which would indicate a potential scale drift between the two equating in the two different time periods.

Method used to detect a difference between two equating functions. The results of the equating function from two different chains were compared both graphically and analytically. Under the graphical approach, the equating functions derived from two different chains were compared using a difference curve. This approach entails calculating the difference between equated scores derived from the two equating chains at each score level and plotting this difference on a graph. Although large differences are often found in the tails of the equating functions because of sparse data, particular attention is given to any difference found in the cut score range. It should be noted that since the difference is calculated for raw-to-scale conversion lines instead of raw-to-raw conversion lines, the difference line from two linear conversions may sometimes appear to be nonlinear. This is because, in any equating, if the old form raw-to-scale conversion is nonlinear, the new form raw-to-scale conversion will be nonlinear, even though the raw-to-raw conversion is linear.

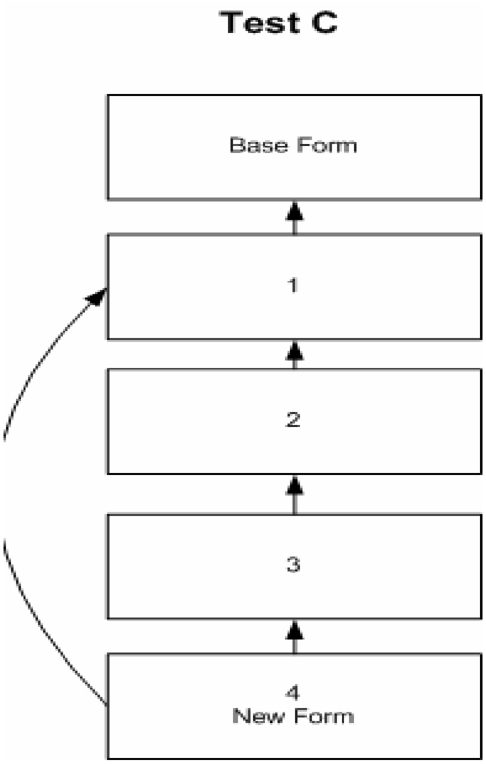

Figure 3. **Single chain equating design for Condition 2.**

Under the analytical approach, equating functions were compared by calculating the root expected squared difference (RESD) between the two equating functions. The RESD is similar to the more conventionally used root mean square difference (RMSD). The only difference between them is that the RESD weights the difference between equated scores at each score level based on the frequency of examinees at that level. The formula for RESD is

$$RESD = \sqrt{\sum_{x=0}^{X} w_x \{[e_1(x) - e_2(x)]^2\}}$$

where x represents each raw score point, e_1 represents the equated scores for a new form via one equating chain, e_2 represents the equated scores for the same new form via another equating chain, and w_x is the weighting factor indicating the proportion of examinees at each raw score level. Smaller values of RESD indicate a negligible difference between two different equating functions.

Criteria used to evaluate a large difference between two equating functions. Dorans and Feigenbaum (1994) proposed the notion of differences that matter (DTM), which stipulates that any difference that is larger than half of the reported score unit should not be considered negligible. Since the tests used in this study use cut scores and scaled scores are reported in 1-point increments, half of this unit can make a difference in the pass/fail status of examinees. For Test A, scores are reported on a scale that extends over 61 scale points, in 1-point increments. For Tests B and C, scores are reported on a scale that extends over 101 scale points, in 1-point increments. Therefore a DTM of 0.5 will be used for the three tests to establish differences between two equating functions (i.e., any difference larger than 0.5 would be considered large because it can change the pass/fail status of examinees).

Pass/fail decisions. Finally, the pass/fail status of examinees based on the different equating functions will be examined. This is important because statistically significant results may not be important practically. On the other hand, if the differences between two equating functions are found to be statistically insignificant but by using either of the equating functions considerable differences in actual pass/fail decisions are observed, then the statistical results may be less useful. This can happen particularly in cases where scores derived using one equating chain as compared to another are very close to each other but would round to different scores. For example a difference between scores such as 50.4 and 50.6 would be undetected using the DTM criteria but 50.4 would round down to 50 and 50.6 would round up to 51. Consequently this may lead to a difference in the pass/fail status of examinees who received such scores. Similarly, often differences larger than the DTM may exist at different score levels but may still not affect the pass/fail status of examinees. For example, a difference between scores such as 50 and 51 is greater than the DTM of 0.5. But if the cut score is 50 then it would not matter whether an examinee received a score of 50 or 51 because both are passing scores.

Results

All equating results presented below were obtained using the chained linear and equipercentile methods under the NEAT design or the direct linear and equipercentile methods under the randomly equivalent group design. The DTM criterion was used for evaluating differences between the equating functions derived from two parallel equating chains or a shorter versus a longer equating chain. Further, difference in the pass/fail rates of examinees using the conversions from these different equating chains was also examined. Results from the first condition (i.e., assessing stability of parallel chains) are presented first followed by results from the second condition (i.e., assessing the stability of a single long chain).

Results for the First Condition (First Design Using Spiraled Forms-Test A)

The means and standard deviations for the new and old forms in this condition are given in Table 1. As seen in Table 1, the mean of the new form is slightly higher than the mean of the two old forms. The SDs of the new form and the second old form are comparable, while the SD of the first old form is slightly higher than the SD of the new form. For the common-item equating (see Figure 2, New Form 7 equated to Old Form 4), the difference in the anchor means divided by the pooled SD of the anchor for the new and old form samples is 0.04, indicating that the ability level of the new and old form samples is about the same.

Results using the graphical approach. The linear and nonlinear equating functions derived via equating Chain 1 and equating Chain 2 are presented in Figures 4 and 5, respectively. As seen in Figure 4, the linear equating functions derived using equating Chain 1 and equating Chain 2 show a larger difference for the lower scaled scores than for the higher scaled scores. Similarly, as seen in Figure 5, the nonlinear equating functions derived using equating Chain 1 and equating Chain 2 also show a larger difference for the lower scaled scores compared to the higher scaled scores.[2]

The difference between the chained linear equating and the direct linear equating via the two different chains for New Form 7 is shown in Figure 6. As seen in the figure, there seems to be a large difference in these two equating functions, especially near the tails of the functions. For this particular test, the cut score range in scaled score units roughly corresponds to about 44 to 58 in the raw score metric. Since users set their cut scores in scaled score units, giving an

Table 1

Summary Statistics for New and Old Forms, All Conditions

	Equating chains	FD	RS	*N*	Mean	SD
Test A: Condition 1,Design 1	Equated via 1st chain using 29 CI	New form	73	3637	51.94	13.57
		Old form	75	4815	49.89	14.20
	Equate via 2nd chain using SG	Old form	72	3640	47.08	13.47
Test B: Condition 1,Design 2	Equated via 1st chain using 12 CI	New form	92	216	68.57	10.29
		Old form	100	95	75.07	9.80
	Equated via 2nd chain using 12 CI	Old form	100	94	71.09	10.71
Test C: Condition 2	Equated via 1st chain using 15 CI	New form	50	1269	24.38	7.98
		Old form	45	2934	22.87	7.23
	Equated via 2nd chain using 15 CI	Old form	50	1591	27.10	7.99

Note. CI = common items; SG = single-group equating design; FD = form designation; RS = total raw score points. New form statistics are same in both chains within each condition and are therefore not presented twice.

exact range of cut scores in raw score units is problematic. Nevertheless a raw cut score range was specified based on how many raw score points are needed to get a particular scaled score for any particular test form. Note that because the raw cut score ranges may vary slightly depending on which equating function (Chain 1 or Chain 2) was chosen, the rough raw score range was determined by taking the lowest and the highest raw cut scores observed across the equatings produced via the two different chains. The same procedure was followed to determine the raw cut score ranges for Tests B and C. As seen from the difference curve, there is a negligible difference (lower than the DTM of 0.5) for the higher cut score regions but not for the lower cut score regions. The difference between the chained equipercentile equating and the direct

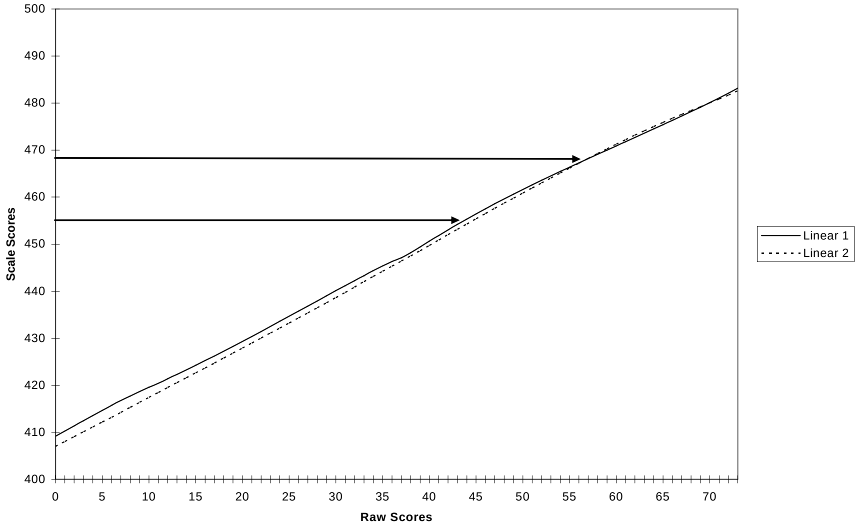

Figure 4. **Condition 1, Design 1 (linear functions derived using two different equatings via separate equating chains).**

Note. The two arrows designate the raw cut score range.

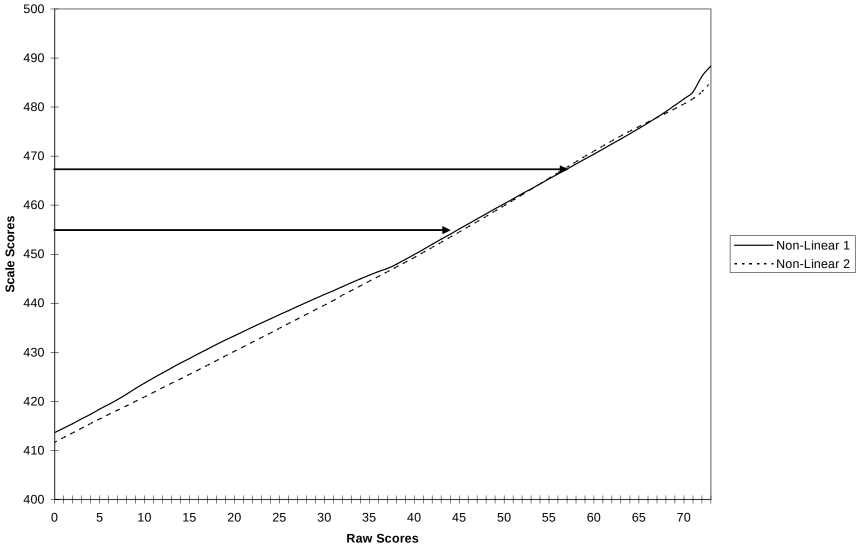

Figure 5. **Condition 1, Design 1 (nonlinear functions derived using two different equatings via separate equating chains).**

Note. The two arrows designate the raw cut score range.

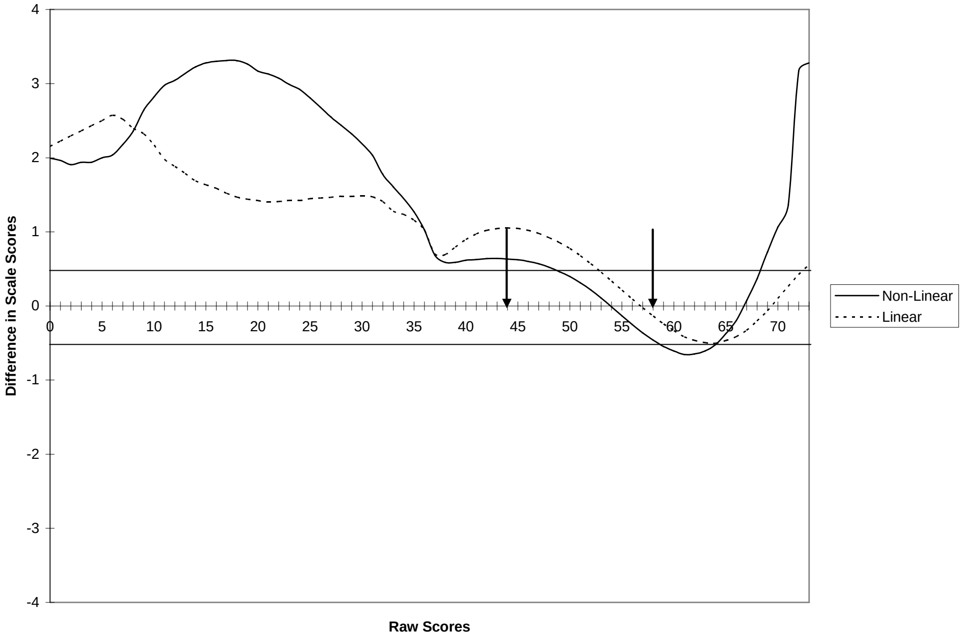

Figure 6. **Difference between equating functions derived using two different equatings via separate equating chains.**

Note. The two arrows designate the raw cut score range.

equipercentile equating via the two different chains for New Form 7 is also shown in Figure 6. As seen in the figure, there seems to be a large difference in these two equating functions especially near the tails of the functions. Also, as in the case of the linear equating functions, there is a negligible difference for the higher cut score regions, but not for the lower cut score regions.

Results using the analytical approach. The RESD values for the above comparisons are presented in Table 2. As seen in this table, the RESD between the linear equating functions for New Form 7 via the two different chains is 0.752. This is larger than the DTM indicating a potentially non-negligible difference between these equating functions. Similarly, the RESD between the equipercentile functions for New Form 7 via the two different chains is 1.094 which is also larger than the DTM, indicating a potentially non-negligible difference between these equating functions. However, because Test A uses cut scores, the difference in the cut score region may still be negligible (i.e., the difference may not make significant differences in the pass/fail status of test takers).

Table 2

RESD Values for Comparisons Between the Equatings From Different Equating Chains

	Equating design	RESD
Condition 1, Design 1	Linear	0.752
	Nonlinear	1.094
Condition 1, Design 2	Linear	2.778
	Nonlinear	3.026
Condition 2	Linear	1.677
	Nonlinear	1.602

Results for the First Condition (Second Design Using NEAT-Test B)

The means and standard deviations for the new and old forms in this condition are given in Table 1. As seen in the table, the mean of the new form is slightly lower than the mean of the two old forms. The SD of the new form is approximately halfway between the SD of the first old form and that of the second old form. For the two common-item equatings, the differences in the anchor means divided by the pooled SD of the anchor for the new and old form samples are 0.05 and -0.08, indicating that the ability level of the new and old form samples are about the same.

Results using the graphical approach. The linear and nonlinear equating functions derived via equating Chain 1 and equating Chain 2 are presented in Figures 7 and 8, respectively. As seen in Figure 7, the linear equating functions derived using equating Chain 1 and equating Chain 2 show a larger difference for the lower and higher scaled score regions, but a small difference in the middle region. Similarly, as seen in Figure 8, the nonlinear equating functions derived using equating Chain 1 and equating Chain 2 show a larger difference for the lower and higher scaled score regions, but a small difference in the middle region.

The difference between the two chained linear equatings via the two different chains for New Form 6 is shown in Figure 9. As seen in the figure, there seems to be a large difference in these two equating functions especially near the tails of the functions. For this particular test, the cut score range in scaled score units corresponds roughly to the 44 to 50 raw score region. As seen from the difference curve, there is a negligible difference (lower than the DTM of 0.5) for the middle and higher cut score regions, but not for the lower cut score regions. The difference

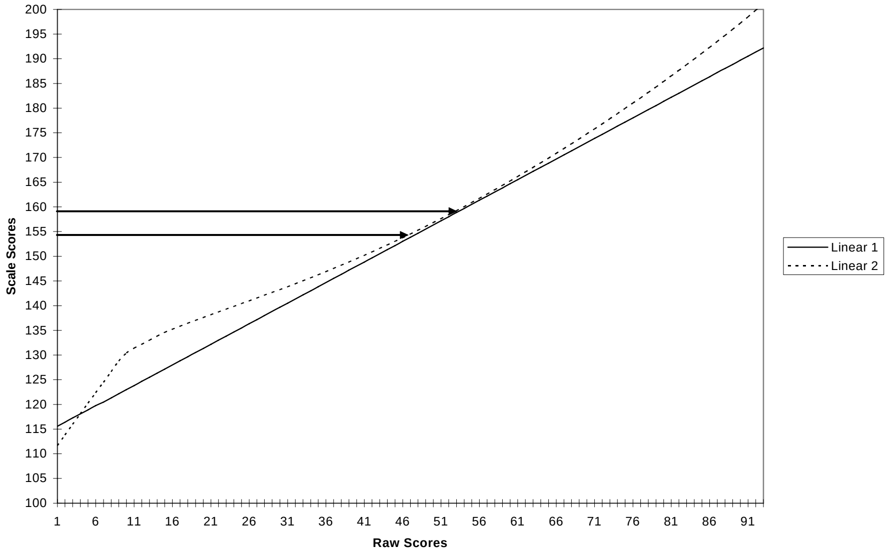

Figure 7. **Condition 1, Design 2 (linear functions derived using two different equatings via separate equating chains).**

Note. The two arrows designate the raw cut score range.

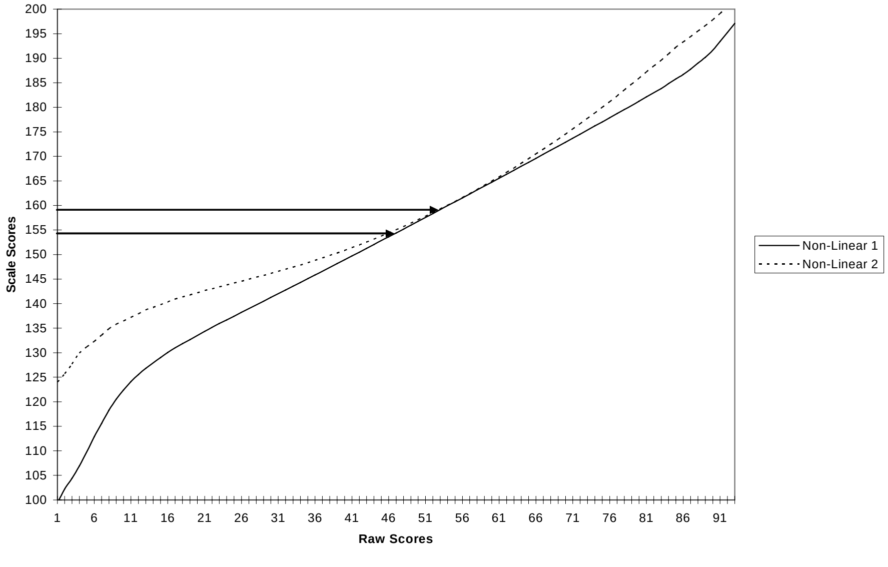

Figure 8. **Condition 1, Design 2 (nonlinear functions derived using two different equatings via separate equating chains).**

Note. The two arrows designate the raw cut score range.

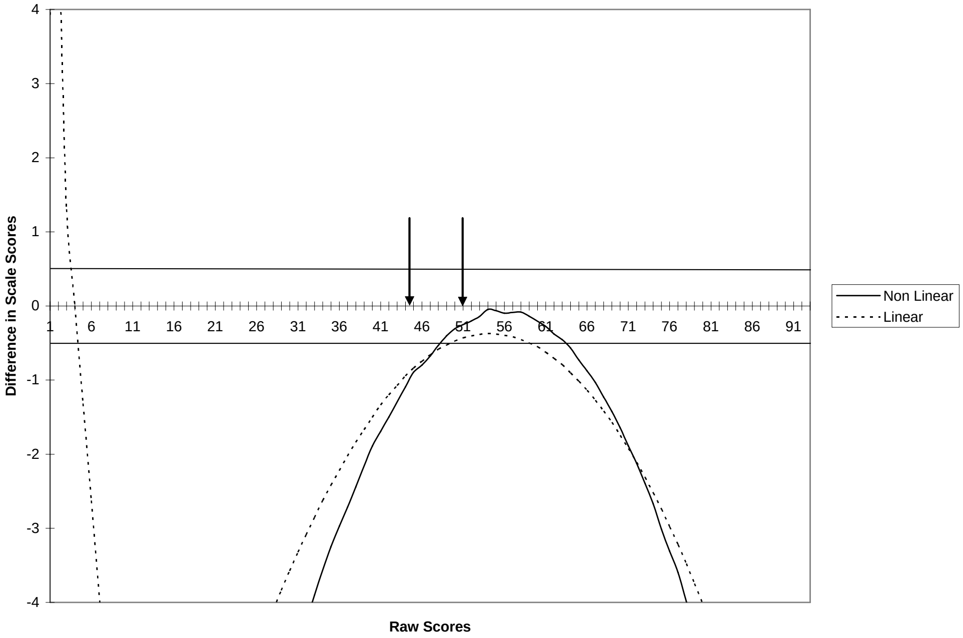

***Figure 9.* Difference between equating functions derived using two different equatings via separate equating chains.**

Note. The two arrows designate the raw cut score range.

between the nonlinear equating functions via the two different chains for New Form 6 is also shown in Figure 9. As seen in Figure 9, there seems to be a large difference in these two equating functions especially near the tails of the functions. Also, as in case of the linear equating functions, there is a negligible difference for the middle and higher cut score regions but not for the lower cut score regions.

Results using the analytical approach. The RESD values for the above comparisons are presented in Table 2. As seen in this table, the RESD between the chained linear functions for New Form 6 via the two different chains is 2.778. This is larger than the DTM, indicating a potentially non-negligible difference between these equating functions. Similarly, the RESD between the equipercentile functions for New Form 6 via the two different chains is 3.026, which is also larger than the DTM and indicates a potentially non-negligible difference between these equating functions. Since Test B also uses cut scores, pass/fail rates will be evaluated to see the actual impact of this difference for test takers.

Results for the Second Condition (Using NEAT-Test C)

The means and standard deviations for the new and old forms in this condition are given in Table 1. As seen in the table, the mean of the new form is slightly higher than the mean of the first old form but slightly lower than the mean of the second old form. The SDs of the new form and the second old form are comparable, while the SD of the first old form is slightly lower than the SD of the new form. For the two common-item equatings, the differences in the anchor means divided by the pooled SD of the anchor for the new and old form samples are 0.02 and -0.07, indicating that the ability level of the new and old form samples are about the same.

Results using the graphical approach. The linear and nonlinear equating functions derived via equating Chain 1 and equating Chain 2 are presented in Figures 10 and 11, respectively. As seen in Figure 10, the linear equating functions derived using equating Chain 1 and equating Chain 2 show a small but potentially non-negligible difference at almost all scale score regions. Similarly, as seen in Figure 11, the nonlinear equating functions derived using equating Chain 1 and equating Chain 2 also show a small but potentially non-negligible difference at almost all scale score regions.

The difference between the two chained linear equatings via the two different chains for New Form 4 is shown in Figure 12. As seen in Figure 12, there seems to be a large difference in these two equating functions especially in the lower half of the functions. The cut score range in scaled score units corresponds roughly to the 17 to 33 raw score region. As seen from the difference curve, there is a potentially non-negligible difference (greater than 0.5) for the total cut score region. The difference between the chained equipercentile equatings via the two different chains for New Form 4 is shown in Figure 12. As seen in Figure 12, there seems to be a large difference in these two equating functions, especially near the lower half and the very top of the upper tail of the functions. Also as in case of the linear equating functions, there is a potentially non-negligible difference for the total cut score region.

Results using the analytical approach. The RESD values for the above comparisons are presented in Table 2. As seen in this table, the RESD between the chained linear functions for New Form 4 via the two different chains is 1.677. This is larger than the DTM indicating a potentially non-negligible difference between these equating functions. Similarly, the RESD between the equipercentile functions for New Form 4 via the two different chains is 1.602, which

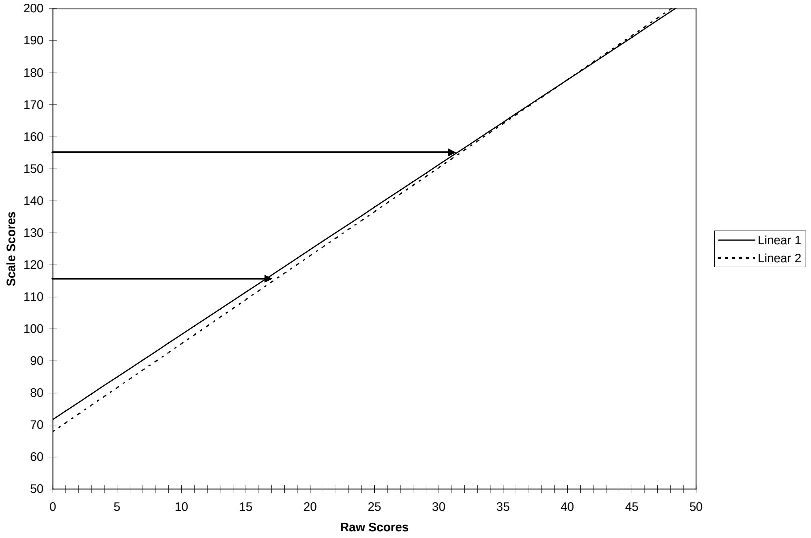

Figure 10. **Condition 2 (linear functions derived using two different equatings via separate equating chains).**

Note. The two arrows designate the raw cut score range.

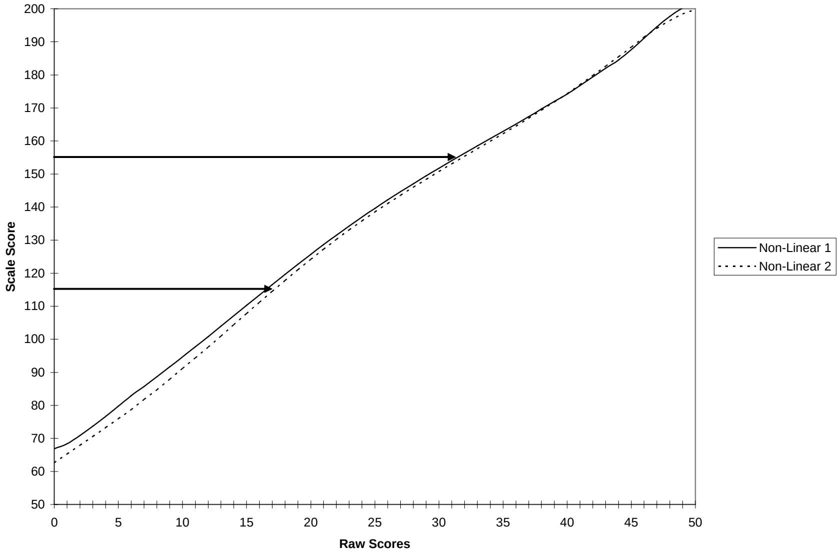

Figure 11. **Condition 2 (nonlinear functions derived using two different equatings via separate equating chains).**

Note. The two arrows designate the raw cut score range.

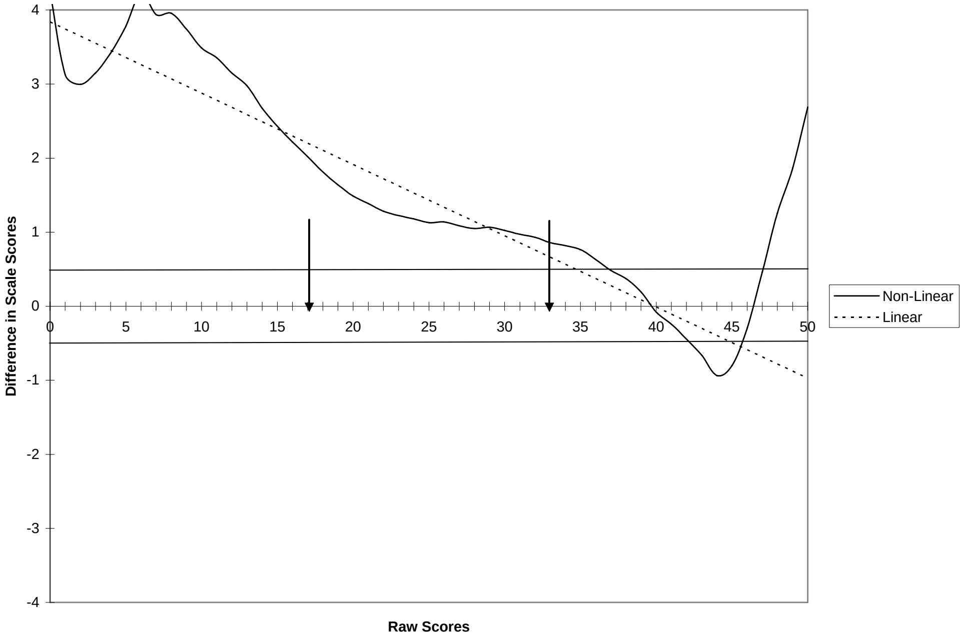

Figure 12. **Difference between equating functions derived using two different equatings via separate equating chains.**

Note. The two arrows designate the raw cut score range.

is also larger than the DTM and indicates a potentially non-negligible difference between these equating functions. Since Test C also uses cut scores, pass/fail rates will be evaluated to see the actual impact of this difference for test takers.

Pass/Fail Results

First condition, first design (using spiraled forms-test A). The actual pass percentages of test takers from different users with different cut scores are reported in Table 3. As seen in Table 3, the pass percentages for test takers using the linear equating functions via the two different equating chains are identical for some users but slightly different for other users using different passing or cut scores. However, the pass percentages for test takers using the nonlinear equating functions via the two equating chains are identical for most users, suggesting a negligible drift between the two equating chains in the cut score region. It should be noted that the number of test takers whose pass/fail status may change could be large or small depending on how many test takers write this test for a particular user state.

Table 3

Actual Pass Percentages of Examinees Using Conversions Derived From Different Equating Chains (Condition 1, Design 1)

User	N	Linear 1	Linear 2	Nonlinear 1	Nonlinear 2
1	104	52.88	52.88	50.00	50.00
2	191	68.59	68.59	67.02	67.02
3	327	**64.83**	**67.58**	64.83	64.83
4	126	**73.02**	**75.50**	70.63	70.63
5	127	**85.04**	**86.61**	83.46	83.46
6	230	**79.57**	**82.61**	78.26	78.26
7	222	76.58	76.58	**74.77**	**76.58**
8	263	**71.48**	**73.76**	71.48	71.48
9	47	76.60	76.60	76.60	76.60
10	355	**68.73**	**70.42**	68.73	68.73
11	33	**72.73**	**78.79**	63.64	63.64
12	262	**79.77**	**80.92**	79.77	79.77
13	47	**63.83**	**68.09**	55.32	55.32
14	79	**58.23**	**60.76**	58.23	58.23
15	96	**65.63**	**66.67**	65.63	65.63
16	100	68.00	68.00	68.00	68.00
17	54	87.04	87.04	87.04	87.04

Note. Users with *N* < 25 are not included in these results. Bold indicates a difference in pass percentage. Users indicate individual states that have adopted the test and specified a cut score.

First condition, second design (using NEAT-test B). The actual pass percentages of test takers from different users with different cut scores are reported in Table 4. As seen in the table, the pass percentages for test takers using the linear equating functions via the two different equating chains are identical for all users, suggesting a negligible drift between the two equating chains in the cut score region. This is also true for the nonlinear equating functions.

Table 4

Actual Pass Percentages of Examinees Using Conversions Derived From Different Equating Chains (Condition 1, Design 2)

User	N	Linear 1	Linear 2	Nonlinear 1	Nonlinear 2
1	39	97.44	97.44	97.44	97.44
2	30	96.67	96.67	96.67	96.67
3	30	96.67	96.67	96.67	96.67
4	100	95.00	95.00	95.00	95.00
5	28	100	100	100	100

Note. Users with *N* < 25 are not included in these results. Users indicate individual states that have adopted the test and specified a cut score

Second condition (using NEAT-test C). The actual pass percentages of test takers from different users with different cut scores are reported in Table 5. As seen in the table, the pass percentages for test takers using the linear equating functions via the two different equating chains are identical for most users. However, the pass percentages for test takers using the nonlinear equating functions via the two different equating chains are identical for some users but slightly different for other users using different passing or cut scores.

Finally, when evaluating scale drift, one may also consider using conditional standard errors of equating, or CSEE (i.e., equating error at each score point), to evaluate differences between two equating functions. For example, if the difference between two scaled scores is 1, but the average CSEE via the two different chains is about 1.5, then one may argue that the difference is within the bounds of sampling error and can therefore be ignored. Note that following a DTM criteria of 0.5, the same difference would be considered large. However, in this study the CSEE approach revealed results similar to those of the DTM approach (see the appendix). For example, in Condition 1, Design 1, the DTM approach showed a negligible difference between the linear equating functions for the higher but not the lower cut score region.

Similarly the CSEE approach showed that the difference between the two linear equating functions fell within the CSEE band for the higher but not the lower cut score region.

Table 5

Actual Pass Percentages of Examinees Using Conversions Derived From Different Equating Chains (Condition 2)

User	N	Linear 1	Linear 2	Nonlinear 1	Nonlinear 2
1	31	**58.06**	**64.52**	**58.06**	**64.52**
2	52	40.38	40.38	**40.38**	**46.15**
3	95	36.84	36.84	43.16	43.16
4	63	66.67	66.67	71.43	71.43
5	31	38.71	38.71	38.71	38.71
6	74	**31.08**	**35.14**	35.14	35.14
7	44	43.18	43.18	**43.18**	**45.45**
8	39	25.64	25.64	**25.64**	**38.46**
9	164	39.02	39.02	**39.02**	**43.29**
10	94	56.38	56.38	**56.38**	**61.70**
11	131	50.38	50.38	56.49	56.49
12	59	23.73	23.73	30.51	30.51
13	104	27.88	27.88	**27.88**	**33.65**
14	35	80.00	80.00	80.00	80.00

Note. Users with $N < 25$ are not included in these results. Bold indicates a difference in pass percentage. Users indicate individual states that have adopted the test and specified a cut score.

Discussion and Conclusions

The purpose of this study was to evaluate scale drift in test equating using data from three tests that employed cut scores. Two commonly observed conditions (i.e., parallel chains and single long chains) in actual testing programs were studied. In the first condition, the new form was equated back to the base scale via two equating chains (i.e., the parallel chains). The conversions derived from the two different chains were compared. Similarly, in the second condition the new form was equated back to the base scale via two equating chains (i.e., new form equated to the previous old form and to another form used some time ago). Again, the conversions derived from the two different chains were compared. If the conversions derived from the parallel and single-chain equating chains are similar, then the pass/fail status of examinees should not be greatly affected using either conversion.

Graphical and analytical approaches were used to evaluate differences between the two equating functions derived via the different equating chains. Overall, although there were substantial differences between equating functions derived from the two different equating chains for the two conditions, the effect of these differences on actual pass/fail status was minimal. Condition 1 (Test A) findings are summarized as follows: (a) The difference curve between the two linear equating functions and the two nonlinear equating functions derived via the two different equating chains showed negligible differences in the higher cut score regions and non-negligible differences in the lower cut score regions; (b) the RESD results obtained by comparing the two linear equating functions and then the two nonlinear equating functions were large (greater than the DTM of 0.5), indicating a non-negligible difference between them; and (c) the pass percentages for the new form sample using the two linear equating functions showed differences in pass rates for some cut scores but not for others. However, the pass percentages using the two nonlinear equating functions showed that the pass rates were identical for almost all cut scores. Using either the graphical or analytical approach, differences were observed between the two equating functions; however, the effect of these differences on actual pass percentages was minimal.

Condition 1 (Test B) findings are summarized next: (a) The difference curve between the linear equating functions and the nonlinear equating functions derived from the two different equating chains showed negligible differences in the middle region of the score scale (i.e., where most of the cut scores are) and non-negligible differences in the tails of the score scale; (b) the RESD results obtained by comparing the two linear equating functions and then the two nonlinear equating functions were large (greater than the DTM of 0.5), indicating a non-negligible difference between them; and (c) the pass percentages for the new form sample using the two linear equating functions showed tno difference in pass rates for all the cut scores. Similarly, the pass percentages using the two nonlinear equating functions showed that the pass rates were identical for all the cut scores. Similar to findings from Condition 1 (Test A), findings from this condition showed that even though differences were observed between the two equating functions using either a graphical or analytical approach, these differences had no effect on actual pass percentages.

Finally, Condition 2 (Test C) findings are summarized as follows: (a) The difference curve between the linear equating functions and the nonlinear equating functions derived from the two different equating chains showed a potentially non-negligible difference for the total cut score region; (b) the RESD results obtained by comparing the two linear equating functions and then the two nonlinear equating functions were large (greater than the DTM of 0.5), indicating a non-negligible difference between them; and (c) the pass percentages for the new form sample using the two linear equating functions showed minor differences in pass rates for most cut scores. However, the pass percentages using the two nonlinear equating functions showed that the pass rates were identical for some cut scores but not for others. Even though differences were observed between the two equating functions using either a graphical or analytical approach, the effect of the differences on actual pass percentages was minimal.

As seen in the findings summarized above, for the two conditions examined, there were some differences in the conversions derived via one chain compared to the other chain. However, the effect of these differences on actual pass/fail status was not large. Larger differences were observed, especially in pass/fail rates, depending on which conversions were compared (i.e., linear or nonlinear). For example, in Condition 1, Design 1, differences were observed for the linear comparisons, but the difference was negligible for the nonlinear comparisons. However, in Condition 2, there were differences observed for the nonlinear comparisons but negligible differences for the linear comparisons.

The reason for this difference in Condition 1, Design 1, is difficult to explain because in this condition, even when the current raw-to-raw equating was linear, the resulting raw-to-scale conversion was nonlinear. Therefore a strict linear versus nonlinear comparison is problematic.[3] For Condition 2, one possible reason for the difference may be that the relationship between the old and new forms was clearly linear and therefore the linear comparisons showed less difference than the nonlinear comparisons. Note that in Condition 2, since a nonlinear conversion was never chosen for any equating in the equating chain, the resulting raw-to-scale conversion for the linear equating results remained linear and therefore could be compared with the nonlinear equating results.

An important question that should be addressed in a study on scale stability is what can be done when scale drift is found, or when it is not found. In the case of the parallel chain condition (Condition 1, Designs 1 and 2), there is no clear answer as to which chain is more accurate. Therefore a reasonable option would be to take the average of the final conversion lines for the two equatings and use their average as the final line for the new form.[4] Another approach would be to examine the standard errors and choose the conversion line that has the lower standard error of equating (although one has to be careful about this because a lower standard error does not always imply a more accurate equating). Similar approaches can also be followed when there is little or no evidence of scale drift between parallel chains as a preemptive measure to check scale drift in test equating.

In the case of the single-chain condition, similar approaches to those prescribed for the parallel-chain condition could be followed. If a difference were found between equating a new form to the previous old form and to another old form administered some time ago, then it may be better to use the equating that went back to the old form administered sometime ago. Doing so would avoid the chain of equatings that can accumulate error. Another option could be averaging the two conversions from both equatings. As a third option, the standard error of equating could be examined to make a judgment regarding which conversion to choose. It should be noted that although the single-chain condition was included in this study, sometimes it is difficult to examine such a condition in actual testing conditions because tests and samples may have evolved considerably over time, making it difficult to find a well functioning anchor set for the new form and an old form administered sometime ago. For tests with large testing volumes, an alternative to this problem would be to identify an old form that was administered sometime ago, making sure that it still meets current testing requirements, and then spiraling this old form with the new form in a current administration. The new form can be equated back to the previous old form using an appropriate equating method (e.g., using a NEAT design). Then the same new form could be equated back to the old form that was administered sometime ago, but spiraled in the current administration, by using the randomly equivalent groups design.

The exact causes of scale instability are difficult to determine because optimal equating conditions are often not completely satisfied in actual testing programs. According to Petersen et al. (1983), there are various factors that may lead to suboptimal conditions for equating. First, the composition of successive groups of examinees who take a particular test may shift over time, leading to highly discrepant ability groups across different time periods. Second, sample sizes available for equating certain tests may be lower than optimal, leading to instability in the equating. Third, tests evolve with time because content areas that are deemed important during one time period may become less important with time. For example, certain laws may become outdated, bringing new laws into force and causing tests to change. As a second example, when calculators were allowed for certain tests for the first time, certain items became useless because they no longer measured what was originally intended (e.g., simple divisions or multiplications). And finally, although equating ideally employs an anchor test that reflects the content and difficulty of the total test, practical constraints often make this impossible. If imperfect anchor tests perpetuate themselves across the equating chain, further instability can result.

Implications for Future Research

The current study did not examine different factors that may cause scale drift. Rather the intent was to identify scale drift (if any) and suggest solutions if drifts were found. Therefore future research could be conducted to systematically examine the effects of factors that affect scale drift. For example, scale drift in equating could be examined relative to differences in ability of new- and old-form test-taking groups. Similarly, scale drift ccould be evaluated relative to the difficulty level, content representation, and actual number of common items used to equate test forms. Finally, this study evaluated scale drift using chained equating methods. Future studies could use other equating methods such as Tucker, Levine, and kernel equating to evaluate whether results from these methods differ from those of the methods used in the current study.

Regardless of the factors that may cause scale drift in equating, it is important to monitor the stability of the reporting scale because, in actual testing programs, new forms are often put on scale through a series of intermediate equatings. This may cause equating error to accumulate to a point where scaled scores are rendered incomparable across two parallel chains or time periods. This is especially important in the case of criterion-referenced tests or tests that employ cut scores, where slight differences can lead to differences in pass/ fail status. However, the findings of this study suggest that although there are differences in the actual conversions derived from the different equating chains, the practical impact of these differences on actual pass/fail status of examinees was minimal.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. A. (1971). *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Donlon, T. F. (1984). *The College Board technical handbook for the Scholastic Aptitude Test.* New York: College Entrance Examination Board.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Guo, F., & Wang, L. (2003, April). *Online calibration and scale stability on a CAT program.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Holland, P. W., & Strawderman, W. (1989). *How to average equating functions if you must.* Unpublished manuscript.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS

McHale, F. J., & Ninneman, A. M. (1994). *The stability of the score scale for the Scholastic Aptitude Test from 1973 to 1984* (ETS Statistical Rep. No. SR-94-27). Princeton, NJ: ETS.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT-verbal score scale* (College Entrance Examination Board Research and Development Report, RDR-74-75, No. 3). New York: College Board.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137–156.

Stewart, E. E. (1966). *The stability of the SAT-verbal scale* (College Entrance Examination Board Research and Development Rep. RDR-66-7, No. 3). New York: College Board.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating.* New York: Springer-Verlag.

Notes

[1] Spiraling refers to distributing two or more test forms in a current administration in alternating sequence. This assures that the groups of test takers taking these test forms are similar in many ways (e.g., test center, test-taking time, knowledge and skills, similar representation of sub-groups, etc)

[2] Although for the two conditions examined in this study, the linear function derived via one chain was compared with the linear function derived via another chain and the nonlinear function derived via one chain was compared with the nonlinear function derived via another chain, it is possible that in certain cases a linear function is defensible for the equating derived through one chain but a nonlinear equating is defensible for the equating derived through another chain. In such cases, it would be more appropriate to compare the equatings that are chosen for the two equating chains (e.g., linear in one chain versus nonlinear in another chain). However, such a case was not observed in the equatings included in the current study.

[3] Some may argue that choosing different conversions (linear and nonlinear) in different equatings in the equating chain clearly limits the interpretation of the results. However, this is often unavoidable, because the choice of a particular equating method is based on criteria that are specific to that equating (e.g., nature of difficulty differences in old and new forms, type of relationship between old and new forms, availability of large data to justify a nonlinear conversion, etc). Therefore, choosing a linear equating for one particular testing administration does not automatically justify choosing a linear equating in subsequent testing administrations. The decision of which equating method to use is usually made based on circumstances specific to that particular testing administration.

[4] A simple averaging of two linear or two nonlinear conversion lines would result in a conversion line that cannot be strictly considered as equating because the resulting conversion line would violate one assumption of equating, namely that equating is reversible. According to Livingston (2004), in practice, the difference between the inverse of an average equating function and the average of the inverses of two separate equating conversions is negligible. Some procedures using angle bisectors have been suggested for averaging equating functions that maintain the reversible property of equating (Holland & Strawderman, 1989).

Appendix

This appendix presents results where the difference between two conversions is evaluated relative to the conditional standard errors of equating (CSEE).

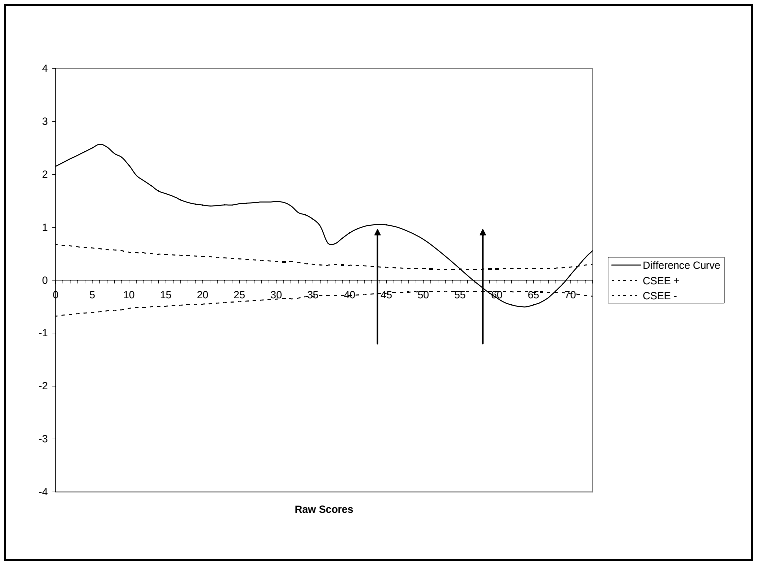

Figure A1. **CSEE and difference between two linear equatings in Condition 1, Design 1.**

Note. The two arrows designate the raw cut score range.

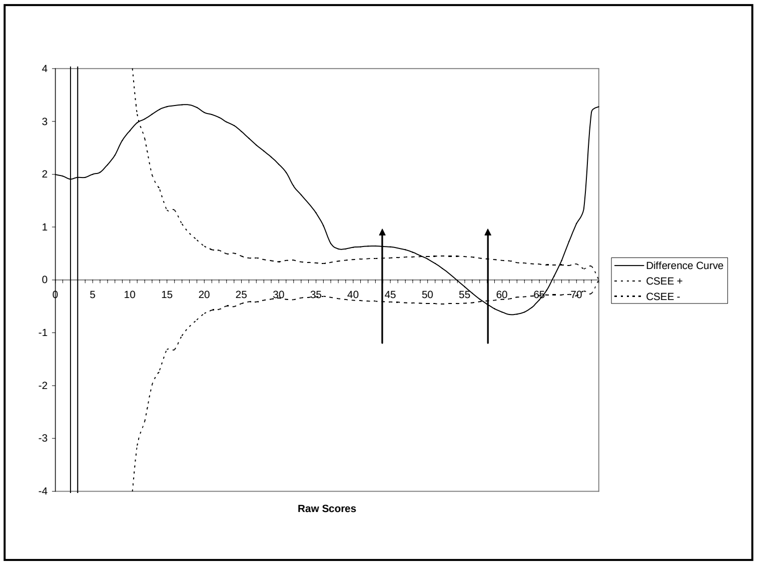

Figure A2. **CSEE and difference between two nonlinear equatings in Condition 1, Design 1.**

Note. The two arrows designate the raw cut score range.

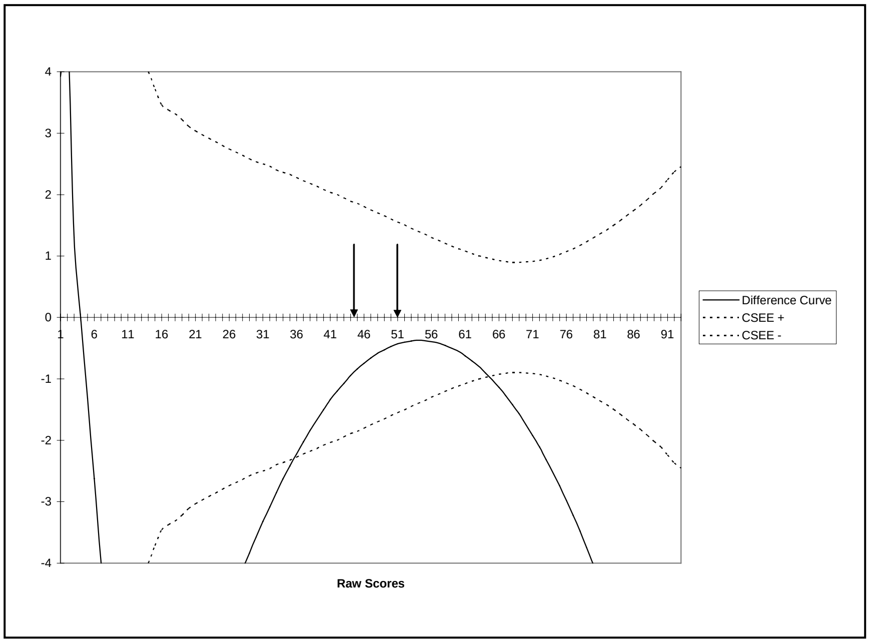

***Figure A3*. CSEE and difference between two linear equatings in Condition 1, Design 2.**

Note. The two arrows designate the raw cut score range.

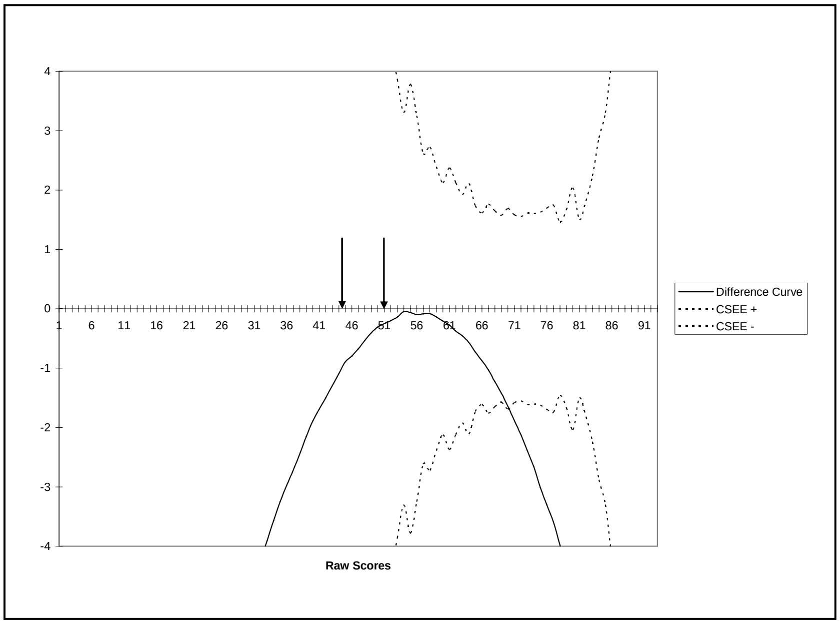

***Figure A4*. CSEE and difference between two nonlinear equatings in Condition 1, Design 2.**

Note. The two arrows designate the raw cut score range.

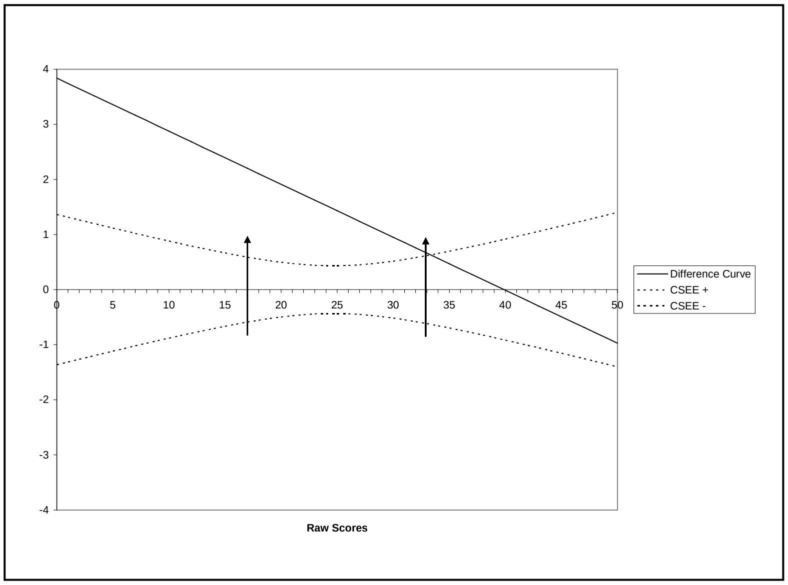

Figure A5. **CSEE and difference between two linear equatings in Condition 2.**

Note. The two arrows designate the raw cut score range.

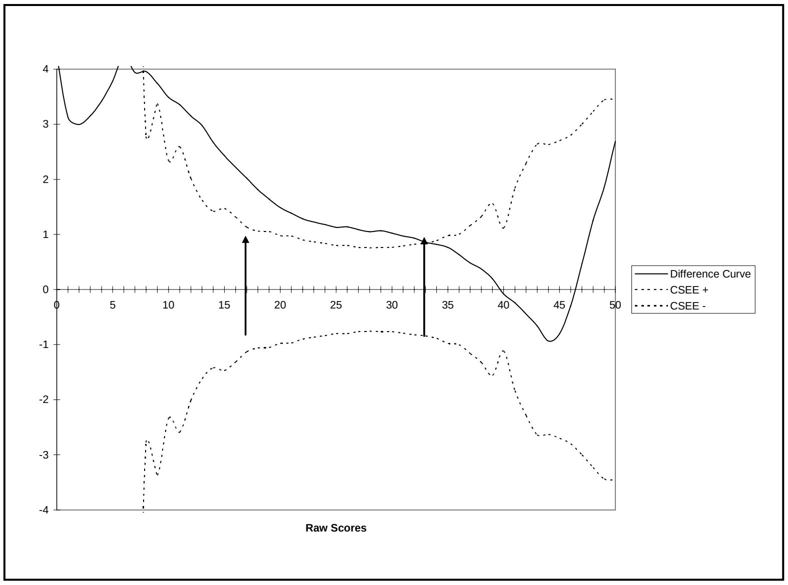

Figure A6. **CSEE and difference between two nonlinear equatings in Condition 2.**

Note. The two arrows designate the raw cut score range. Instead of using the average standard errors from two equatings, one may also use the standard error of equating difference (SEED), which is the standard error of the difference between two equating functions (see von Davier, Holland, & Thayer, 2004, chap. 5).